Prediction of Contiguous Regions in the Amniote Ancestral Genome

Aïda Ouangraoua, Frédéric Boyer, Andrew Mcpherson, Eric Tannier, Cedric Chauve

Prediction of Contiguous Regions in the Amniote Ancestral Genome

Aïda Ouangraoua[1], Frédéric Boyer[2], Andrew McPherson[1], Éric Tannier[3], and Cedric Chauve[1]

[1] Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada
aouangra|awm3|cchauve@sfu.ca
[2] Institut de Recherches en Technologies et Sciences pour le Vivant; Laboratoire Biologie, Informatique et Mathématiques; CEA Grenoble, F-38000 Grenoble, France
frederic.boyer@cea.fr
[3] INRIA Rhône-Alpes; Université de Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622, Villeurbanne, France
Eric.Tannier@inria.fr.

Abstract. We investigate the problem of inferring contiguous ancestral regions (CARs) of the genome of the last common ancestor of all extant amniotes, based on the currently sequenced and assembled amniote genomes as ingroups and three teleost fish genomes as outgroups. We combine a methodological framework using conserved syntenies computed from whole genome alignments of amniote species together with double conserved syntenies (DCS) using gene families from amniote and fish genomes, to take into account the whole genome duplication that occurred in the teleost lineage. From these comparisons, ancestral genome segments are computed using techniques inspired by physical mapping. Due to the difficulty caused by the whole genome duplication and the large evolutionary distance to the closest assembled outgroup, very few methods have been published with a reconstruction of the amniote ancestral genome. This one is the first which is founded on a simple and formal methodological framework, whose good stability is shown and whose CARs cover large regions of the human and chicken genomes.

To appear in the Proceedings of ISBRA 2009, 5th International Symposium on Bioinformatics Research and Applications, May 13-May 16, 2009, Nova Southeastern University, Fort Lauderdale, Florida, USA. Version of February 18, 2009.

1 Introduction

The reconstruction of ancestral karyotypes and gene orders from homologies between extant species can help to understand the large-scale evolutionary mutations that differentiate the present genomes. It has been approached using cytogenetics methods and recently applied to mammalian genomes [24]. Beyond this evolutionary distance, homologies are less visible and it is only with the recent availability of sequenced and assembled genomes that bioinformatics

methods can predict the past of chromosomes. These methods address the problem at a much higher resolution, although with much less available genomes. The first results have been obtained on mammalian genomes [4, 20, 16], and several reviews have been published [9, 19], analyzing the divergences with earlier cytogenetics results [10, 5, 22]. These methods can be divided into model-based methods, that compute complete evolutionary scenarios [4, 20] and model-free approaches that do not consider a precise rearrangement model, which are used by cytogeneticians and currently receive a lot of attention from computational biology (see [16, 6] and references there).

The application of such methods to more ancient genomes comes up against the difficulty to handle duplications and losses as evolutionary events. Yet the teleost fish genomes have undergone a whole genome duplication (WGD) [12, 19] at an early stage of their evolution, and are currently the only available genomes that may serve as an outgroup to reconstruct amniote or tetrapod ancestral genomes. Two recent methods have been developed to reconstruct these ancestral genomes [21, 14], and predict very divergent syntenic associations. Hence, while the reconstruction of the ancestral mammalian genome seems now to be close to a relative consensus, the reconstruction of the amniote ancestral genome looks as the next bottleneck on the way towards the ancestral proto-vertebrate genome.

In [6], a general model-free framework was introduced for the reconstruction of "Contiguous Ancestral Regions", or CARs (the terminology is borrowed from Ma *et al.* [16]) in an ancestral genome. It is inspired by genome physical mapping techniques, and roughly consists in two phases as follows. Given a set of "genomic markers", which are sets of orthologous positions in the ingroup genomes: (1) detect "ancestral syntenies", which are sets of genomic markers that are believed to have been contiguous in the ancestral genome, and (2) order the genomic markers into a set of "Contiguous Ancestral Regions" in which the ancestral syntenies are respected, discarding some of them if the whole set is not compatible with the formation of linear CARs. This second phase relies on combinatorial tools such as PQ-trees, that were introduced in computational biology for physical mapping of genomes. Indeed, our problem consists in the mapping of markers into ancestral chromosomes. This framework was applied in [6] for the reconstruction of contiguous ancestral regions of mammalian ancestors (ferungulates and boreoeutheria). It was shown to be very stable under different parameters for the computation of syntenies.

Our goal here is to apply this framework to compute CARs of the ancestral amniote genome. The method used to infer mammalian CARs needs to be extended to handle two main issues. First, the closest currently available sequenced and assembled outgroups are the teleost fishes, whose evolutionary distance to the ingroups (mammals and birds) is considerable. Hence it is impossible to define a high coverage set of genomic markers that appear once in each genome of this study. Moreover, the Whole Genome Duplication followed by massive gene losses and intensive rearrangements makes ancestral syntenies inaccessible by a classical comparison between amniote and fish genomes.

We handle these issues by using (1) genomic markers obtained from whole-genome alignments within amniote assembled genomes (chicken and mammals), and (2) gene families to compare amniote and fish genomes (teleost fishes) to construct ancestral syntenies. We rely on the Double Conserved Synteny (DCS) principle introduced by Kellis *et al.* [13] and Dietrich *et al.* [7], since then often used to detect syntenies in a WGD context [12, 21], and systematized by Van de Peer [23]. Despite the principle is well known, the detection of such syntenies implies a non trivial methodological problem and formal descriptions of the expected signal, adapted to the highly rearranged fish genomes, are lacking. It is a contribution of this paper to propose a formal definition of DCSs.

We obtain a family of ancestral syntenies, and group them into Contiguous Ancestral Regions of the proto-amniote genome. This set can contain some "conflicting signal", which means that no linear ordering of the genomic markers can account for all the ancestral syntenies. While the ancestral syntenies computed from mammalian genomes presented very little conflict [6], the greater evolutionary distance mixes up the signal and we have to cope with more conflicting ancestral syntenies. That is why we produce here the sets of CARs with different sets of parameters used to compute the DCS and propose both a set of CARs obtained with stringent parameters and a set of consensus CARs obtained from several values of parameters. We compare our results with two other studies [14, 21] that proposed a configuration of the amniote ancestral genome. These two present contradictory results, low coverage of the extant genomes by the reconstructed ancestral one, and no validation of the methods. We try to make significant progresses in these directions: our CARs are more numerous than in the previous studies but present a good coverage of the extant genomes. As in the previous studies, we find a proto-amniote genome that shows more similarity to the chicken genome than to mammalian genomes.

In the following we first describe the method, following the framework of [6], and focusing on the novelty we introduce, which is the possibility of integrating duplicated syntenies in this framework. The definition and computation of the duplicated syntenies is discussed. Then we describe the CARs we obtain into details, comparing them to previous studies, showing some convergences and differences. Eventually we study the soundness of the proposed CARs by running the method under different sets of parameters, to understand its behavior, its stability and the confidence we may have on the proposed CARs.

2 Data and methods

2.1 Overview

We consider a dataset containing eleven amniote genomes (human, chimpanzee, orangutan, macaca, mouse, rat, dog, cow, horse, opossum, chicken) and three teleost fish genomes (tetraodon, stickleback, medaka) used as outgroups. We then proceed with the following steps.

1. We compute a set of "genomic markers" that are unique and universal in amniote genomes (i.e. each markers appears once and exactly once in each of the eleven genomes), using whole genome alignments available.
2. A first set of ancestral syntenies is generated by computing common intervals (as in [6]) of genomic markers between all pairs of amniote species whose evolutionary path goes through the amniote ancestor (here, chicken against every mammal).
3. A second set of ancestral syntenies is generated by computing "double conserved syntenies" (DCS) between each amniote and the three teleosts, using sets of gene families. This provides the coordinates on an amniote genome of a genomic segment which is likely to descend from a segment of the ancestral osteichthyes genome (the ancestor of teleost fishes and amniotes), and is then likely to have been present as an segment of the ancestral amniote genome. These coordinates provide a set of genomic markers (those which intersect the DCS) which is taken as the ancestral synteny.
4. We weight the ancestral syntenies from both sets according to their conservation pattern in the considered species tree.
5. We select from the ancestral syntenies a maximum weight subset such that alltogether they are compatible with the formation of linear CARs. This supposes that the least weight conflicting ancestral syntenies are more likely to be false positives. This phase relies on the "consecutive ones problem", widely used in physical mapping problems. The final result is a combinatorial structure (a PQ-tree) that allows to linearly represent the whole set of solutions to the consecutive ones problem. The children of the root of the PQ-tree are the amniote CARs.

2.2 Genomes and markers

We used data retrieved from the Compara database of Ensembl [8], comparing human (hg18), chimpanzee (CHIMP2.1), orangutan (PPYG2), macaca (Mmul_1) mouse (mm9), rat (RGSC 3.4), dog (CanFam2.0), cow (Btau_4.0), opossum (monDom5), platypus (Ornitorhynchus_anatinus-5.0), chicken (WASHUC2), tetraodon (TETRAODON8.0), medaka (HdrR) and stickleback (BROAD S1). For the phylogenetic relationships between the amniote species, we considered the species tree used by Compara to compute whole-genome alignments (see `http://www.cecm.sfu.ca/~cchauve/SUPP/ISBRA09/` for data and results).

We construct a set of genomic markers by using the Pecan 12-amniotes-vertebrates multiple alignments available in the release 50 of Compara with human genome as a reference. From the orthologous seeds on the 11 fully assembled genomes defined by the multiple alignments, we keep only the ones that have a minimum size (100b in the human reference genome). Two seeds are joined if they are distant from each other by less than 100Kb in all amniotes where they occur. A *genomic marker* is then an inclusionwise maximal set of linked seeds which spans more than 100Kb in all genomes and such that its seeds span at least 50% of its total span. In this way we expect to obtain a good set of orthologous markers, removing uncertain homologies as well as paralogies. We

obtain 1101 non-overlapping genomic markers, spanning respectively 797Mb of the human genome (26% of its size) and 308Mb (29%) of the chicken genome.

2.3 Ancestral syntenies.

As described in Chauve *et al.* [6], ancestral synteny detection in the absence of large duplications (in mammalian and bird genomes) may be performed by the detection of groups of genomic markers contiguous in at least two genomes whose evolutionary path goes through the desired ancestor. But the WGD that is believed to have occurred in the lineage of the teleost genomes requires a more sophisticated treatment when a fish genome is involved in the comparison.

In the absence of large duplications: comparing two amniote genomes. First, when comparing the chicken genome to each mammalian genome, which are not separated by a whole genome duplication event, an *ancestral synteny* is the set of genomic markers intersecting (a) an inclusionwise maximal segment of the chicken genome that contains the same genomic markers as a segment of a mammalian genome (a *maximal common interval*) or (b) a segment of the chicken genome that contains only two markers that are also consecutive in a mammalian genome (an *adjacency*). We include adjacencies to balance the fact that there is no order information associated with common intervals.

The result of this process is a family of sets of markers, each of which covers amniote genomic segments that are believed to be ancestral.

In the presence of a WGD in the outgroups. For the comparisons between an amniote genome and a fish genome, the method described above can not be applied due to the numerous losses and intra-chromosomal rearrangements that shuffled the teleost fish chromosomes after the WGD. Indeed, massive losses of genes often follow a WGD, and two paralogous segments may present little similarity. It is possible to detect this paralogy indirectly by comparing the two segments with their common ortholog in a non duplicated genome. This principle is called the *pivot* method, and is now classical to detect chromosome segment homologies in a WGD context [13, 7, 12, 23] (see Figure 1). Despite this principle is well known, no methodological discussion on the exact signal that should be detected using fish genomes has been published (due to the highly rearranged fish chromosomes, the approaches described in [23] are not efficient in this case). So in the present method we propose a definition of a DCS. The proportion of conflicting signal in the whole set of DCS tends to show that there is still some space for improvement of the precision of this proposition.

We use a set of orthologous gene families constructed from the orthologies between genes of amniotes and fishes available in the Ensembl Compara database [8][4]. When comparing an amniote genome to a fish genome, we use only genes that are annotated with coordinates on a chromosome on both species. A

[4] We use release 51, which is based on the same Genome assemblies than release 50, with improved gene annotation, but does not contain Pecan alignments.

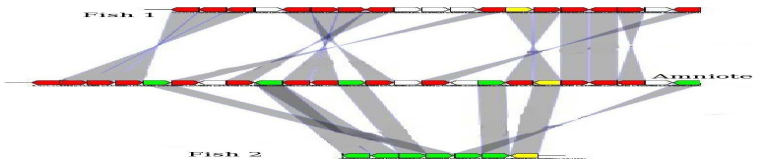

Fig. 1. A double synteny: an amniote chromosome (in the middle) is homologous to two fish segments (up and down), though paralogy between these two segments is not detectable through direct similarity (few genes are conserved in two copies). Clusters are much rearranged: this justifies the method of detection, which does not consider the order of the genes within a cluster.

double conserved synteny (DCS) is a segment S of the amniote genome that is orthologous to two paralogous segments S_1 and S_2 in two different chromosomes of the fish genome, *i.e.* satisfies the two following criteria.

1. S contains at least 20 genes, min_{prop} percent of all genes having orthologs on S_1 or S_2 (we choose $min_{prop} = 95\%$ and test the sensitivity to this parameter),
2. There is a minimum number of alternances (4 for this study) along the genes of S between those which have an ortholog on S_1 and those which have an ortholog on S_2.

These conditions, inspired by the comparison between tetraodon and human genomes in [12], were designed to retrieve genome segments whose gene content exhibits a clear signal for originating from a single genome segment pre-WGD. We considered here all comparisons between one of the three teleost genomes and each amniote. We obtained a list of amniote genome segments orthologous to two segments in two chromosomes of a fish, showing a signal for a "double synteny". The maximal set of genomic markers intersecting such a segment defines the corresponding ancestral synteny.

We also included ancestral syntenies from mammalian ancestral segments showing a DCS signal with fishes. These are sets of genomic markers intersecting maximal common intervals between two mammalian genomes whose evolutionary path goes through the boroeutherian ancestor, which are in addition included in a DCS in at least one of the two considered genomes. These segments are refined DCS, with a stronger conservation signal, as they are predicted to be ancestral by two methods, in the boreoeutherian ancestral genome and in the osteichtyes ancestral genome.

We obtain 2745 ancestral syntenies containing more than one genomic marker. These syntenies are then weighted according to the pattern of conservation they present in the phylogeny, using the formula described in [6] that accounts for a species tree and the branch lengths.

2.4 Assembling ancestral syntenies

The output of the phase described above is a set $\mathcal{L}$ of n genomic markers, and a family $\mathcal{S} = \{S_1, \ldots, S_m\}$ of m subsets of $\mathcal{L}$, where each subset is a set of genomic markers that are believed to be contiguous in the ancestral genome of interest. Following [6], we use the approach traditionally applied to physical mapping problems [1]. It is based on the *consecutive ones property* (C1P) and *PQ-trees*. We encode $\mathcal{S}$ by an $m \times n$ 0/1 matrix $\mathcal{M}$ where row i represents S_i as follows: $\mathcal{M}[i, j] = 1$ if marker j belongs to S_i and 0 otherwise. Ordering markers into CARs consists in finding a permutation of the columns of the matrix $\mathcal{M}$, such that all 1's entries in each row are consecutive (also called a C1P ordering for $\mathcal{M}$). Finding such an order of the columns of $\mathcal{M}$ is not always possible, in particular if there are false positives in $\mathcal{S}$, that is groups of markers that were not contiguous in the ancestral genome. Moreover, if there exists a C1P ordering of the columns of $\mathcal{M}$, there are often several possible orderings that make all 1's consecutive on each row, that represent different ancestral genome architectures.

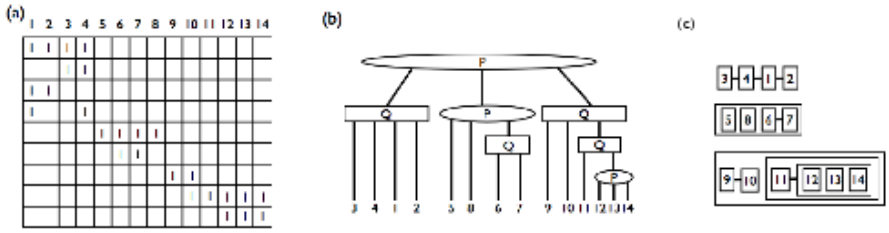

Fig. 2. (a) A matrix $\mathcal{M}$ with the consecutive ones property. (b) A PQ-tree $T(\mathcal{M})$. (c) An equivalent representation of $T(\mathcal{M})$ that highlights all ancestral genome architectures that correspond to C1P orderings for $\mathcal{M}$: each row corresponds to a chromosomal segment represented by a child of the root, two glued blocks have to be adjacent in any ancestral genome architecture and sets blocks that float in the same box have to be consecutive in any genome architecture but their order is not constrained. Here for example we see three ancestral chromosomal segments and the second one contains markers 5 to 8, with only constraint that markers 6 and 7 are adjacent; hence, 5 6 7 8 is a possible order for this last segment, but not 5 6 8 7. All 13824 possible C1P orderings (possible ancestral orderings) are visible on this representation, that we use to present the amniote CARs in Figure 3.

If $\mathcal{M}$ is not C1P, then we know that some sets of markers in $\mathcal{S}$ are false positives and were not contiguous in the ancestral genome. Following [16, 6], we clear ambiguities by computing a maximal subset of $\mathcal{S}$ that is C1P, using a branch-and-bound algorithm described in [6] that finds an exact solution. Then, given this C1P subset of $\mathcal{S}$, all C1P orderings can be represented in a compact way, using the *PQ-tree* of the resulting matrix $\mathcal{M}'$, denoted $T(\mathcal{M}')$, that contains three

kinds of nodes: leaves (labeled by $\mathcal{L}$), P-nodes and Q-nodes. Computing $T(\mathcal{M}')$ can be done efficiently [17]. See Figure 2 for an illustration. $T(\mathcal{M}')$ encodes in a compact way all possible C1P orderings of the columns of $\mathcal{M}$ and then all genome architectures we can deduce from $\mathcal{S}$: the root of $T(\mathcal{M})$ is a P-node, the children of the root represent CARs, where Q-nodes describe fixed orderings, up to a reversal, while P-nodes, but the root, describe subsets of markers that have to be contiguous but where there is no information to fix a relative order (see Figure 2 for an illustration). Two markers that are consecutive children of a Q-node are said to define an *adjacency*.

3 Results and discussion

3.1 An ancestral amniote genome architecture

We applied the described method to propose a genomic architecture of the ancestral species of all amniotes, using the data presented in the previous section. Of the 2745 ancestral syntenies, 372 had to be removed during the optimization phase, in order to have some solutions to the C1P problem. This resulted in an ancestral amniote genome architecture composed of 79 CARs, 63 of them containing more than one genomic marker. In these 63 CARs, 983 of 1101 genomic markers are included in adjacencies, which indicates that there is little ambiguity due to P-nodes in the PQ-tree, although more than in the boreoeutherian CARS of [6]. These CARs define 282 segments that are strictly colinear with segments of the chicken genome, and cover 75% of the chicken genome. Similarly, these CARs define 225 segments that are colinear with the human genome and cover 67% of this genome. Although these numbers are smaller than for reconstruction of mammalian ancestors described in [6], they are much larger than the amniote CARs inferred in [21]. The 51 CARs which span more than 1 Mb of the chicken genome are illustrated on Fig. 3, with their correspondence with the chicken chromosomes.

Kohn *et al* [14]	Nakatani *et al.* [21]	Present method
2-9-16, 1-24, 5-10,	2-9	2-12, 22-Z
17-Z, 4-22, **18-27**-19,	13-**17-Z**,	**17-Z**, **18-27**
21-26-23-32, 3-14, 8-18	1-7, 1-14-18	**21-26**, 1-8, 4-20, 9-19

Table 1. Recovered syntenic associations between chicken chromosomes, for three different methods. Numbers refer to pieces of chicken chromosomes. CARs which contain markers from only one chicken chromosome are not mentioned here.

Chicken syntenic associations. A "syntenic association" is the presence in a single CAR of genomic markers from two different chromosomes of an extant amniote. Here we may observe several chicken syntenic associations and compare

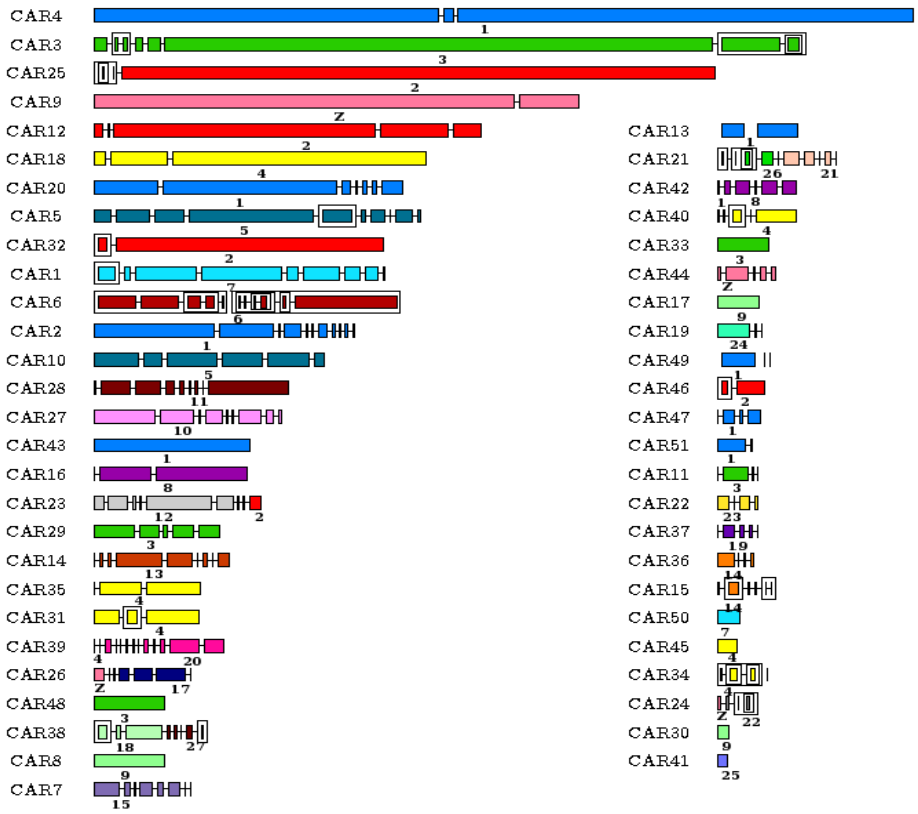

Fig. 3. The PQ-tree of amniote CARs, with their correspondence in the chicken genome. All ancestral architectures are represented, in the format described in Figure 2, with few chromosomic segments in which the order of the markers is not fixed.

them with other published methods [14, 21]. These are, up to our knowledge, the only two methods that lead to the proposition of an architecture for the amniote genome. These propositions are very divergent: not only the number of chromosomes varies between 18 [14] and 26 [21], but the observed syntenic associations between chicken chromosomes are not always compatible. Of 13 syntenic associations found by Kohn *et al.* [14] and 6 found by Nakatani *et al.* [21], only two are common (Z-17 and 2-9). The reason is probably the absence of a formal framework, that we tend to fill here. We give a summary of the differences in Table 1, together with the syntenic associations we find in this study. We find one of the common syntenic associations (17-Z), plus two associations from Kohn *et al.* [14] and none additional from Nakatani *et al.* [21].

3.2 Stability of the method and sensitivity to parameters

The advantage of using a general framework for ancestral genome reconstruction is the possibility, to a certain extent, of assessing the quality and robustness of the results. We have a simple support to every pair of adjacent markers in the CARs: the existence of an ancestral synteny that contains these markers. So every adjacency in CARs may be examined using the data, independently of the methodology.

The choices we have to make are the parameters used in the computations of ancestral syntenies. When two amniotes are compared, the ancestral synteny construction requires no parameter and its stability has been assessed in [6].

When an amniote genome and a fish genome are compared, we rely on a novel method that is less tried and tested, that is the computation of DCS. The principle itself is well known and employed, but we are not aware of any formal study on a reliable implementation of this principle. The two parameters that are used to construct the DCS are inclusive: strengthening both parameters will improve the specificity. Any DCS which is found for a set of parameters will also be found by less stringent parameters. The question is which parameters are stringent enough to assure a good specificity. We think an amniote segment with at least 20 genes, covering 95% of the genes annotated in this segments is a sufficient proof for an double orthology signal, and it is confirmed by the comparison with the map of [12] made from the same principle with visual expertise. The optimization step is also a source of possible instability of the method, especially as, from our experiments with the stringent criterion $min_{prop} = 95$, at least 10% of the possible ancestral syntenies are false positive or result from convergent evolution, and have to be discarded during the optimization phase.

To assess these two possible sources of instability, we inferred DCS with the following values of the parameter min_{prop}: 80%, 83%, 86%, 89%, 92% and 95%. To measure the stability, we considered the set of adjacencies (pairs of markers that are consecutive children of a Q-node) defined by each set of CARs. An adjacency is said to be conserved between two sets of CARs if it is found in both sets of CARs. An adjacency of a given set of CARs is said to be weakly conserved in another set of CARs if it is absent in the latter one but the two markers that define it belong to the same CAR. We show in Table 2 the characteristics of the set of CARs we computed.

positive ancestral syntenies, as the ratio of discarded syntenies drops from 29% to 14%. However, the results obtained are very stable in terms of adjacencies, as most adjacencies of a given set of CARs are consistent with the adjacencies in the previous set of CARs. The increase in the number of CARs is expected as less ancestral syntenies are obtained with more stringent parameters to compute DCS. This seems to indicate that, on this particular dataset, the optimization phase behaves very consistently with various sets of DCS and seems to conserve a subset of DCS that lead to very similar sets of CARs. Note also that the proportion of genomic markers that do not belong to adjacencies (i.e. are children of a P-node) is low, which indicates that the proposed ancestral

min_{prop}	Ancestral Syntenies	Discarded Syntenies	CARs	Long CARs	Adjacencies	Conserved Adj.	Weakly conserved Adj.
80	4054	1182	25	20	1062	-	
83	3691	973	37	24	1044	1021	17
86	3328	779	47	34	1033	1012	15
89	3093	626	49	35	1034	1011	13
92	2956	491	70	46	1006	979	24
95	2745	372	79	63	983	961	16

Table 2. Characteristics of the set of CARs computed with DCS obtained with several values of min_{prop}. Discarded syntenies are the ancestral syntenies discarded during the optimization phase. Long CARs are CARs that cover at least 1Mb of the chicken genome. Conserved and weakly conserved adjacencies are in terms of the previous value of min_{prop}.

genomic architectures contain very few segments where the order of the markers is not fixed.

Additional results regarding the support of each adjacencies by the different types of ancestral syntenies – common intervals between ingroups, DCS and mammalian syntenies included in DCS – are available on the companion website. They show that with these data, that contain only one species on one branch from the ancestor (birds and reptiles), DCS are fundamental to detect and support a significant number of adjacencies: 112 adjacencies are not supported by any ancestral synteny chicken-mammalian.

4 Conclusion

We extended a general framework for reconstructing ancestral genome architectures [6] in order to handle WGD events. We apply our method to reconstruct the architecture of the ancestral amniote genome. While we put a lot of attention to the specificity of the method, not to infer doubtful ancestral syntenies, sensitivity is not sufficient to provide the exact set of chromosomes of the amniote ancestor. The small number of genomes used, as well as the fact that the chicken is the only available genome among birds and reptiles and has several very small chromosomes, is also a reason for which there are certainly more CARs than ancestral chromosomes. The definitive amniote ancestral genome is still an open problem, but with this general, simple and formal method, some of its characteristics are accessible and valuable for further studies.

From a methodological point of view, several avenues can be explored. One of the issues is the fact that the C1P framework requires that each genomic marker appears exactly once in the ancestral genome architecture. This forced us to define our genomic markers from whole-genome alignments. But already at the level of the amniotes, these markers span around 30% of the chicken genome. Another approach could have been to use genes as genomic markers. However, in order to apply the C1P framework, this requires to compute the gene content

of the ancestral amniote genome using the gene trees/species tree reconciliation approach. This problem is still a hard problem, that is very sensitive for example to errors in computing gene trees [11]. The other main issue is the computation of the DCS. The results we obtain with several values of the parameter that defines DCS clearly show that many of the DCS we compute are probably not ancestral syntenies. They are very likely to intersect or contain genome segments that really originate from an ancestral amniote genome segment, but due to the lack of flexibility of the C1P framework, they are considered as false positive. Indeed, an amniote segment that only overlaps, even on a large part, a complete segment derived from the amniote common ancestor, will induce conflict with respect to the C1P framework. In the same time, without DCS, a significant number of adjacencies are not supported, which makes DCS instrumental in our method. This motivates the problem, open up to now, of the design and study of formal methods for the reliable and precise detection of DCS.

Acknowledgments

Aïda Ouangraoua is funded by the ANR BRASERO (ANR-06-BLANC-0045) and SFU. Eric Tannier is funded by the ANR (ANR-08-GENO-003-01 and NT05-3_45205) and by the CNRS. Cedric Chauve is funded by NSERC and SFU.

References

1. F. Alizadeh *et al.* Physical mapping of chromosomes using unique probes. *J. Comp. Biol.*, 2:159–184, 1995.
2. A. Bergeron, C. Chauve and Y. Gingras. Formal models of gene clusters, in: A. Zelikovsky and I. Mandoiu (eds) *Bioinformatics Algorithms: Techniques and Applications*, Wiley Interscience, Wiley Series on Bioinformatics, pp. 177–202.
3. K.S. Booth and G.S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. System Sci.*, 13:335–379, 1976.
4. G. Bourque, P.A. Pevzner, and G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse and rat genomes. *Genome Res.*, 14:507–516, 2004.
5. G. Bourque, G. Tesler, and P.A. Pevzner. The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome Res.*, 16:311–313, 2006.
6. C. Chauve and E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genome. *PLoS Comput. Biol.* 4:e1000234, 2008
7. F.S. Dietrich *et al.* The *ashbya gossypii* genome as a tool for mapping the ancient *saccharomyces cerevisiae* genome. *Science*, 304:304–307, 2004.
8. T. J. P. Hubbard *et al.* Ensembl 2007. *Nucl. Acid. Res.*, 35:D610–D617, 2007.
9. T. Faraut. Addressing chromosome evolution in the whole-genome sequence era. *Chromosome Res.*, 16:5–16, 2008.
10. L. Froenicke *et al.* Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome Res.*, 16:306–310, 2006.

11. M. W. Hahn. Bias in phylogenetic tree reconciliation methods: implications for verte- brate genome evolution. *Genome Biol.*, 8:R141, 2007.
12. O. Jaillon *et al.* Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431:946–957, 2004.
13. M. Kellis, B.W. Birren, and E. S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *saccharomyces cerevisiae*. *Nature*, 428:617–624, 2004.
14. M. Kohn *et al.* Reconstruction of a 450-my-old ancestral vertebrate protokaryotype. *Trends Genet.*, 22:203–210, 2006.
15. J. Ma *et al.* DUPCAR: Reconstructing contiguous ancestral regions with duplications *J. Comput. Biol.* 15:1007–1027, 2008.
16. J. Ma *et al.* D. Haussler, and W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, 16:1557–1565, 2006.
17. R.M. McConnell. A certifying algorithm for the consecutive-ones property. In *SODA 2004*, pages 761–770, 2004.
18. J. Meidanis, O. Porto, and G.P. Telles. On the consecutive ones property. *Discrete Appl. Math.*, 88:325–354, 1998.
19. M. Muffato and H. Roest Crollius. Paleogenomics, or the recovery of lost genomes from the mist of times. *BioEssays*, 30:122–134, 2008.
20. W.J. Murphy *et al.* Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309:613–617, 2005.
21. Y. Nakatani, H. Takeda, and S. Morishita. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.*, 17:1254–1265, 2007.
22. M. Rocchi, N. Archidiacono, and R. Stanyon. Ancestral genome reconstruction: An integrated, multi-disciplinary approach is needed. *Gen. Res.*, 16:1441, 2006.
23. Y. Van de Peer. Computational approaches to unveiling ancient genome duplications. *Nat. Rev.*, 5:752–763, 2004.
24. J. Wienberg. The evolution of eutherian chromosomes. *Curr. Opin. Genet. and Dev.*, 14:657–666, 2004.